BREAST CANCER DETECTION

(Course Name: Introduction to Python Programming Lab)

(Course Code: 20CS3352)

A Python Project Report on Breast Cancer Detection

Submitted by

G.V.V.LALITH KRISHNA(22501A0563)

A.GREESHWANTH(22501A05I3)

B.NAVYA SRI(22501A0517)

B.SOWMYA(22501A0524)

CH.SRUTHI(23505A0502)

II B.Tech I Sem

in

Computer Science and Engineering

Prasad V Potluri Siddhartha Institute of Technology

Accredited with A+ grade by NAAC, NBA Accredited,

and Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

Kanuru, Vijayawada-520007

Prasad V Potluri Siddhartha Institute of Technology

Accredited with A+ grade by NAAC, NBA Accredited,

and Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

Kanuru, Vijayawada-520007

CERTIFICATE

This is to certify that the python project report titled "**Breast Cancer Detection"** of **Mr. G.v.v.Lalith Krishna (22501A0563) Mr.A.Greeswanth (22501A0513), Miss. B. Navya Sri (22501A0517), Miss. B. Sowmya(22501A0524), Miss.CH.Sruthi (23505A0502)** in Computer Science and Engineering during the academic year 2023-2024.

Signature of the Guide **Signature of the H.O.D**

TABLE OF CONTENTS:

TITLE	PAGE NOS
Abstract	4
1. Introduction	5
1.1 Background	5
1.2 Motivation	6
1.3 Problem Statement	6
2. Literature Review	7
3. Methodology	8
3.1 Data Collection	8
3.2 Data Pre-Processing	8
3.3 Data Visualization	10
4. Project Design	11
4.1 Data Flow Diagram.	11
5. Implementation	12
5.1 Algorithms Used	12
5.2 Code Development	12
6. Results and Analysis	13
6.1 Performance Evaluation Metrics	14
6.2 Results	15
7. Conclusion	16
8. Achievements and future work.	17

Abstract:

Global cancer data confirms more than 2 million women diagnosed with breast cancer each year reflecting majority of new cancer cases and related deaths, making it significant public health concern. But fortunately, it is also the curable cancer in its early stage. Early diagnosis of breast cancer with timely and effective treatment services improves the prognosis and survival of patients. Duringclassifyingtumors,therearesignificantchancesoferrorandfalse diagnosis which is needed to be worked upon. Accurate classification can prevent patients from unnecessary treatments. Thus, it is important to accurately classify patients into malignant and benign groups with right diagnosis. This study is based on machine learning (ML) algorithms, aiming to review python technique and its application in breast cancer diagnosis and prognosis by building simple machine learning model. Machine learning has unique advantage as it detects critical features from complex breast cancer datasets. The methodology is widely used for classification of pattern and forecast modelling. The primary data for this study is extracted from Wisconsin breast cancer database (WBCD). It is the benchmark database which compares result via different algorithms.

Breast cancer is one of the leading cause for the death of women. In women Breast cancer is treated as the most significant issue. According to statistics released by the International Agency for Research on Cancer (IARC) in December 2020, Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide. Early diagnosis of this helps to prevent the cancer. If breast cancer is detected in early stage, then Survival rate is very high. Machine Learning methods are effective ways to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis for decision making. In this paper, Data Visualization and performance comparisons between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), K Nearest Neighbours (k-NN), Adaboost, XGboost and Random Forest conducted on Wisconsin breast cancer Dataset. The main objective is to evaluate the accuracy in the classification of data in terms of efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Our aim is to review various Techniques To detect early, efficiently and accurately Using Machine Learning. Experimental results show that XGboost offers the highest accuracy (98.24%) with the lowest error rate.

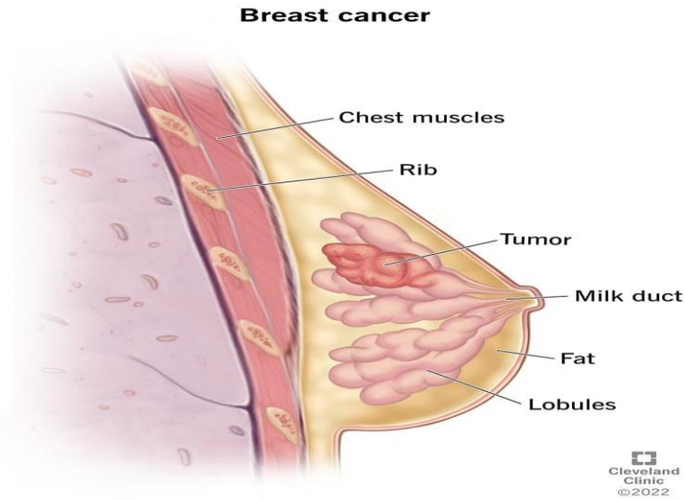
INTRODUCTION:

1.1 Background:

Breast cancer is a significant global health concern, affecting millions of women and, in rare cases, men. According to the World Health Organization (WHO), breast cancer is the most common cancer among women worldwide, both in developed and developing countries. The impact of this disease on individuals and communities is profound, necessitating continuous efforts to enhance early detection, diagnosis, and treatment.

The prevalence of breast cancer underscores the urgent need for effective detection strategies. Statistics from reputable health organizations reveal alarming trends, with a rising incidence of breast cancer cases in various age groups.Current breast cancer detection methods primarily rely on mammography, clinical breast exams, and self-examinations. While these approaches have been valuable in identifying potential cases, they are not without limitations. Advancements in technology present an opportunity to revolutionize breast cancer detection.. This project aims to contribute to the ongoing efforts in breast cancer detection by exploring innovative methods that harness the power of data analytics, machine learning algorithms, and advanced imaging technologies.

Fig-1: Breast with cancer:



1.2:Motivation:

Motivation for breast cancer detection is rooted in the desire to improve early diagnosis, treatment outcomes, and overall survival rates for individuals at risk of or affected by breast cancer. Several key factors drive the emphasis on early detection:

Improved Survival Rates: Early detection allows for timely intervention and treatment, significantly improving the chances of survival. The earlier breast cancer is diagnosed, the more treatment options are available, and the better the prognosis.

Quality of Life: Detecting breast cancer at an early stage often results in less aggressive treatment, reducing the physical and emotional impact on individuals. This can contribute to a better overall quality of life during and after treatment.

Reduced Treatment Costs: Early detection may lead to less invasive and less costly treatments. By identifying breast cancer in its early stages, healthcare systems can potentially reduce the financial burden associated with extensive and prolonged treatments.

Personal Empowerment: Regular breast cancer screening empowers individuals by providing them with information about their health. Awareness and early detection enable people to actively participate in their healthcare decisions and take steps to manage their health proactively.

Public Health Impact: Breast cancer is a significant public health concern globally. Early detection programs and awareness campaigns contribute to a healthier population by reducing the overall burden of advanced-stage breast cancer cases.

Support for Research: Early detection initiatives contribute valuable data to research efforts focused on understanding the biology of breast cancer, identifying risk factors, and developing more effective treatments. The more we learn through early detection, the closer we come to finding a cure for breast cancer.

In summary, the motivation for breast cancer detection is multifaceted, encompassing individual well-being, public health, advancements in medical technology, and the collective effort to reduce the impact of breast cancer on society. Early detection remains a cornerstone in the fight against breast cancer, offering the potential for improved outcomes and a better quality of life for those affected.

1.3:Problem Statement:

Given a dataset of patients with various physiological, biochemical, and clinical features, develop a machine learning model that can accurately and efficiently predict the presence Breast cancer . The model should also provide insights into the most important factors that influence the outcome of Breast cancer and suggest possible interventions to prevent or delay its progression. The model should be validated on unseen data andcompared with existing methods and benchmarks. The model should also be interpretable and explainable, and adhere to ethical and privacy standards.

2.Literature Review:

Relevant Previous Work:

Many works have been submitted which attempted to diagnose breast cancer using machine learning algorithms. For instance, Sun at al. in year 2005 [15], proposed comparing feature selection methods for a unified detection of breast cancers in mammograms. Another approach, introduced by Malekatal in year 2009 [16],proposed a method using wavelet and proposed a design of automated detection, segmentation, and classification of breast cancer nuclei using a fuzzy logic for feature exaction and classification respectively. Zheg at al. in year 2014 [17] combined support vector machine (SVM) and K-means algorithm for breast cancer diagnosis. Aličković and Subasi in year 2017 [18] applied a genetic algorithm for feature extraction and rotation for classification. Another approach is conducted by Bannaiein year 2018 [19] based on the dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) technique to attain output of interest. There are several other works performed based on clustering and classification [20]. Alireza Osarech, Bita Shadgar achieved 98.80% and 96.63% accuracies upon using SVM classification technique on two different benchmark datasets for breast cancer [21]. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza applied KNN, SVM, Gaussian Naïve Bayes, and Logistic Regression techiniques programmed in MATLAB to diagnose and predict recurrence of breast cancer.

The classification technique were applied on two dataset from UCI depository. One dataset was used for identification of diseases (WBCD), and other is used for prediction of recurrence [22]. Vikas Chaurasia, BB Tiwari and Saurabh Pal build predictive models on breast cancer and compared their accuracies using famous algorithms viz. J48, Naïve Bayes, and RBF. The results indicated that Naïve Bayes predicted well among them with 97.36% accuracy [23]. Haifeng Wang and Sang Won Yoon developed a powerful model for breast cancer prediction by using and comparing Naive Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree and Artificial Neural Networks (ANN). They implemented PCA for dimensionality reduction [24]. S.Kharya proposed Artificial Neural Networks (ANN) while working on breast cancer prediction. The paper highlighted advantages of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees [25]. Naresh Khuriwal and Nidhi Mishra used Wisconsin Breast Cancer database to work on breast cancer diagnosis. Based on their experiments they concluded that, ANN and Logistic Algorithm worked better and achievedan accuracy of 98.50% [26].

3.Methodology:

The methodology aims to analyse the most helpful feature in prediction of malignant and benign tumor. This may help to visualize general trend in selecting appropriate model. The objective is to classify benign and malignant tumors of breast cancer with the help of python. The focus is on using Logistic Regression.

We have obtained Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle. Here 569 Patient's Data Was used for analysis, each instances have 32 Attributes with Diagnosis and Features. Each instance has a parameter of the cancerous non-cancerous cells and we will predict the cancer just by the input of features. The values of features is in Numeric Format. The 'Target' means the patient Who is having Whether 'Benign' or 'Malignant' Cancer state. Benign means the patient is not having Cancer and Malignant means the patient is having Cancer.

3.1 Data Collection:

In this paper, Wisconsin Breast Cancer Diagnostics (WBCD) dataset is used which isobtained from UCI Machine Learni0ng Repository [14]. It was created by Dr. William H.Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.The data consists of 569 patients and 32 characteristics. These characteristics formed 32columns in the dataset. Ten highlights of these characteristics are as per following:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave points
- Symmetry
- Fractal dimension- the mean, standard error, and worst (mean of the three largest values) of all the features or characteristics.

3.2:Data Pre-Processing:

Data Preprocessing is a crucial step preceding model building, designed to eliminate unwanted noise and outliers from the dataset, ensuring optimal model training. The process begins after acquiring the relevant dataset, focusing on cleaning the data and preparing it for subsequent model development. The chosen dataset comprises 569 instances and 32 attributes, as outlined in the table.

There are 569 rows and 33 columns which represent 569 patients with 33 data points or features for individual patient in this dataset.

When training a machine learning model on such imbalanced data, it is important to be cautious of metrics beyond accuracy. While accuracy may be high, metrics like precision and recall become more insightful. Given the relatively minor discrepancy between instances suggesting breast cancer and those indicating its absence, there might be no immediate need for undersampling the dataset.

The data contains a column ‘id diagnosis’ in which the input is given as normal for a patient with no breast cancer and input ‘breast cancer’ for a patient with glaucoma. The calculation of
the prediction of breast cancer result are string values and this line of code replaces the entries with the string to a machine understandable and predictable for logistic regression and for linear regression we have converted them as normal to ‘0’ and breast cancer to ‘1’. The values with the string are set as integer values.

Input:

```
data.info
```

```
data.duplicated()
```

Output:

```
0      False
1      False
2      False
3      False
4      False
       ...
564    False
565    False
566    False
567    False
568    False
Length: 569, dtype: bool
```

3.3 Data Visualization:

Data visualization is the discipline to understand data by placing it into visual form in order to interactively and efficiently convey insights so that the patterns, trends and correlations of the data that might not otherwise be detected can be visualized in large data sets. It removes the noise from the data and highlights the useful information.

As visualization makes it easier to detect patterns, trends and outliers, and provides clear, better and reliable result, it is implemented in this paper by creating Line plot and Histogram.

Histogram..

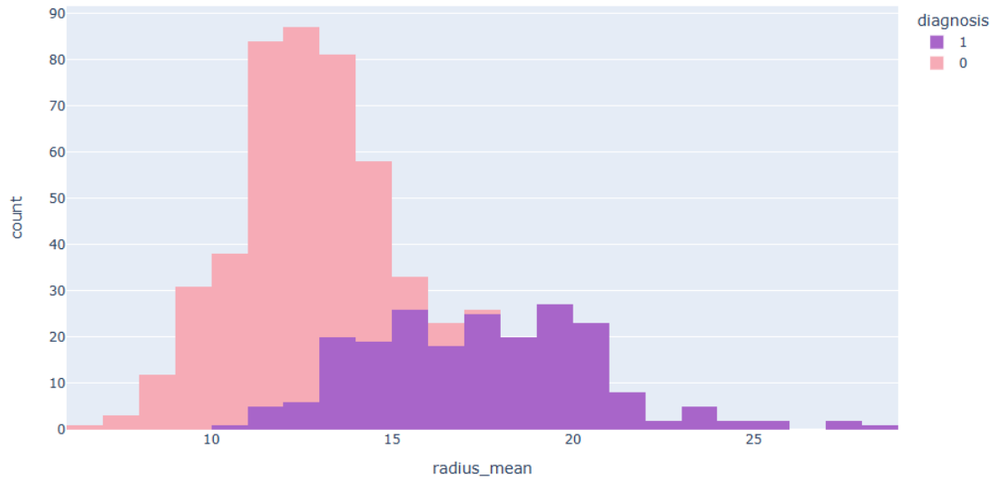


Fig-Showing representation for the persons whom wants diagnosis..

Line plot:

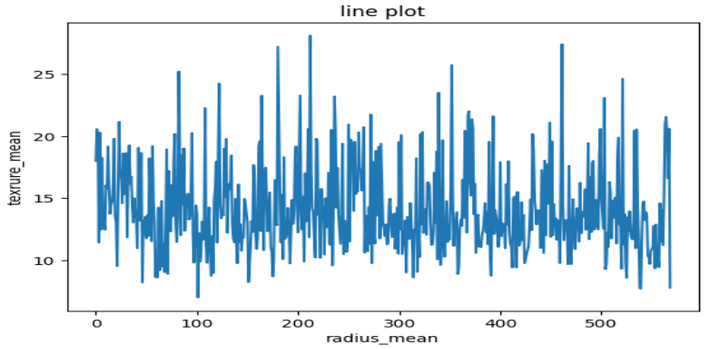
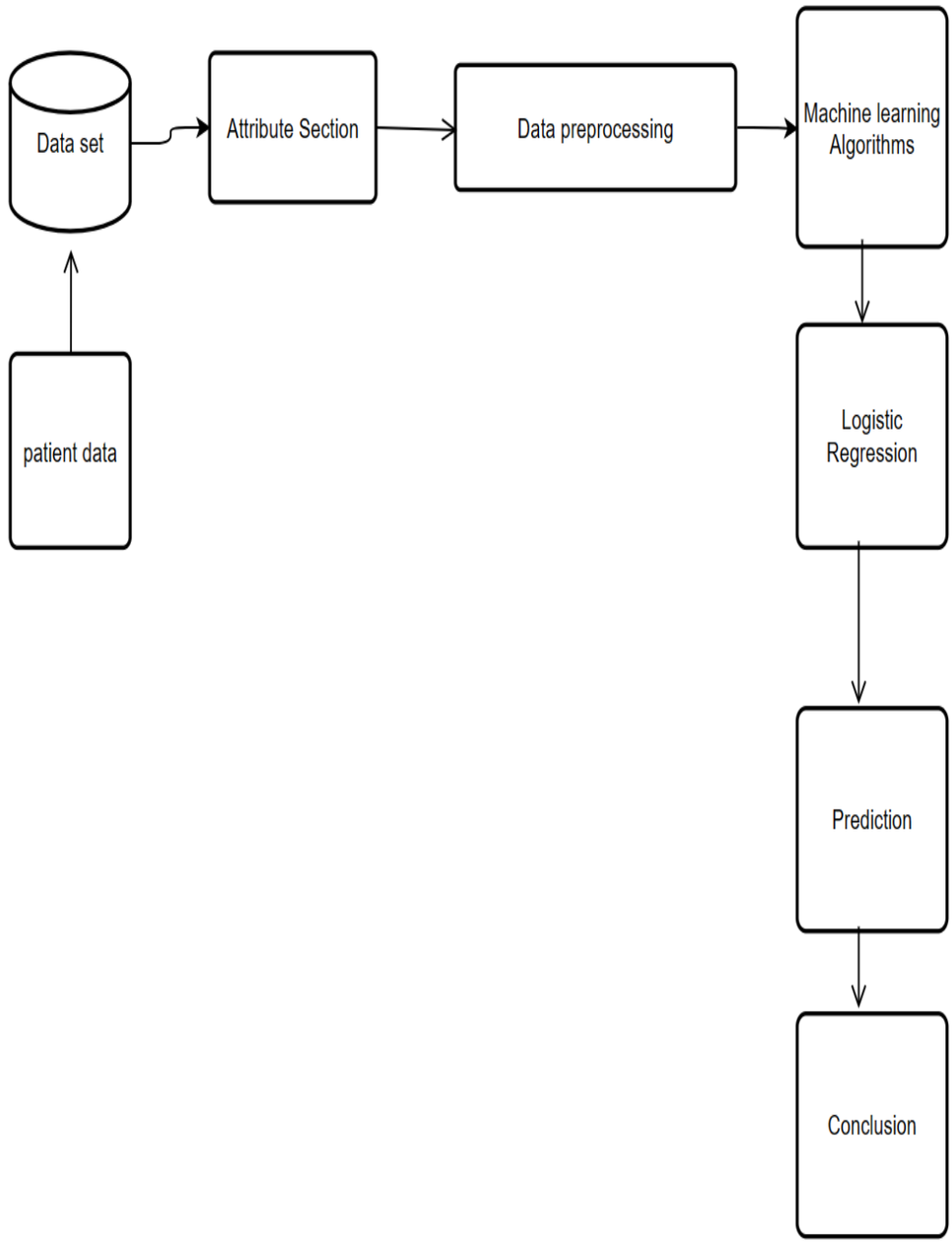


Fig-showing the line plot for radius_ mean, texture_ mean

4.Project Design:

4.1 Data Flow Diagram:



5. Implementation:

5.1 Algorithms Used:

Logistic regression:

Logistic regression is used for predicting the probability of an event occurring (binary outcome) based on one or more predictor variables. It's commonly employed in classification problems, such as spam detection or medical diagnosis.The logistic function (sigmoid function) is used to transform the linear combination of predictor variables into probabilities between 0 and 1.Logistic regression coefficients represent the change in the log-odds of the dependent variable for a one-unit change in the independent variable. Maximum Likelihood Estimation (MLE) is used to find the optimal coefficients in logistic regression. Odds ratios can be derived from logistic regression coefficients to quantify the impact of independent variables on the odds of the event occurring. The confusion matrix is used to evaluate the performance of a logistic regression model in classification tasks.

5.2 Code Development

Data representation and pre processing:

```
[ ] from sklearn.preprocessing import LabelEncoder
    target='diagnosis'
    label_encoder=LabelEncoder()
    data[target]=label_encoder.fit_transform(data[target])
```

```
[ ] import matplotlib.pyplot as plt
    plt.plot(data["radius_mean"])
    plt.xlabel("radius_mean")
    plt.ylabel("texrure_mean")
    plt.title("line plot")
    plt.show()
```

```
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report

# Assuming 'classification' is a variable containing the target column name
classification = 'diagnosis'  # Replace with your actual target column name

# Select features (X) and target variable (y)
feature_columns = ['id',
 'diagnosis',
 'radius_mean',
 'texture_mean',
 'perimeter_mean',
 'area_mean',
 'smoothness_mean',
 'compactness_mean',
 'concavity_mean',
 'concave points_mean',
 'symmetry_mean',
 'fractal_dimension_mean',
 'radius_se',
 'texture_se',
```

```
    'area_se',
    'smoothness_se',
    'compactness_se',
    'concavity_se',
    'concave points_se',
    'symmetry_se',
    'fractal_dimension_se',
    'radius_worst',
    'texture_worst',
    'perimeter_worst',
    'area_worst',
    'smoothness_worst',
    'compactness_worst',
    'concavity_worst',
    'concave points_worst',
    'symmetry_worst',
    'fractal_dimension_worst']
X = data[feature_columns]
y = data[classification]

# Replace '\t?' with NaN
X.replace('\t?', np.nan, inplace=True)
```

```
# Convert columns to numeric (assuming that they are numeric features)
X = X.apply(pd.to_numeric, errors='coerce')

# Impute missing values using the mean strategy
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

# Split the data into training and testing sets
train_X, test_X, train_Y, test_Y = train_test_split(X_imputed, y, test_size=0.1, random_state=50)

# Initialize and train the Logistic Regression model
model = LogisticRegression()
model.fit(train_X, train_Y)

# Make predictions on the test set
predictions = model.predict(test_X)

# Evaluate the model
accuracy = metrics.accuracy_score(predictions, test_Y)
print('The accuracy of the Logistic Regression model is:', accuracy)

# Display the classification report
report = classification_report(test_Y, predictions)
print("Classification Report:\n", report)
```

Accuracy:

```
The accuracy of the Logistic Regression model is: 0.7368421052631579
Classification Report:
              precision    recall  f1-score   support

           0       0.74      1.00      0.85        42
           1       0.00      0.00      0.00        15

    accuracy                           0.74        57
   macro avg       0.37      0.50      0.42        57
weighted avg       0.54      0.74      0.63        57
```

6. Results and Analysis:

6.1Performance Evaluation metrics

When evaluating a machine learning model for predicting Breast cancer, we typically use various performance metrics to assess its effectiveness. Below are some common performance metrics for Breast cancer prediction:

Accuracy:

Accuracy remains a measure of the overall correctness of predictions, representing the ratio of correctly predicted instances to the total number of instances. However, caution should be exercised with accuracy, especially in the context of imbalanced breast cancer detection data.

Precision:

Precision in the breast cancer detection project is the ratio of true positive predictions to the total number of positive predictions made. It signifies how accurately the model identifies cases of breast cancer.

Recall (Sensitivity or True Positive Rate):

Recall measures the model's ability to identify all actual cases of breast cancer by calculating the ratio of true positive predictions to the total number of actual breast cancer cases.

F1-Score:

The F1-Score, in the context of breast cancer detection, acts as the harmonic mean of precision and recall. It provides a balanced measure, crucial when optimizing the trade-off between false positives and false negatives.

Specificity (True Negative Rate):

Specificity gauges the model's ability to correctly identify instances without breast cancer. It is calculated as the ratio of true negative predictions to the total number of actual non-breast cancer cases.

Area Under the ROC Curve (AUC-ROC):

The ROC curve illustrates the model's trade-off between true positive rate (recall) and false positive rate at various thresholds. AUC-ROC quantifies the model's ability to distinguish between breast cancer and non-breast cancer cases.

Area Under the Precision-Recall Curve (AUC-PR):

The Precision-Recall curve, specific to breast cancer detection, plots precision against recall at different thresholds. AUC-PR provides insights into the precision-recall trade-off.

Confusion Matrix:

The confusion matrix, relevant to breast cancer detection, offers a detailed breakdown of true positives, true negatives, false positives, and false negatives. It provides a comprehensive view of the model's performance.

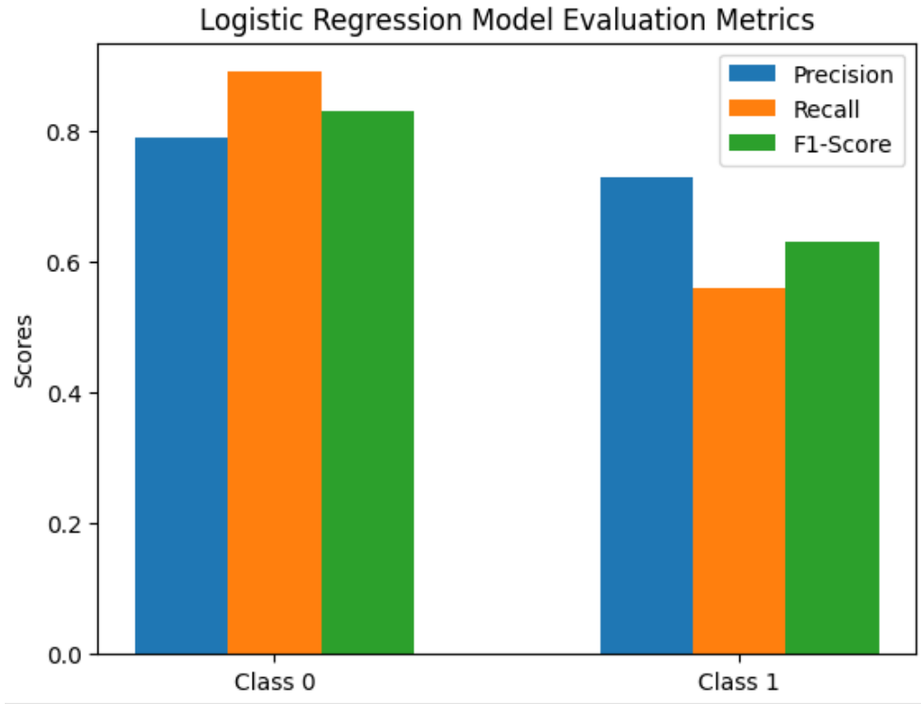
False Positive Rate (FPR):

The FPR in breast cancer detection represents the ratio of false positive predictions to the total number of actual non-breast cancer cases. It indicates the model's tendency to incorrectly predict breast cancer.

True Negative Rate (TNR):
TNR, synonymous with specificity in this context, measures the model's ability to correctly identify instances without breast cancer.

Evaluation Metrics:

```
import matplotlib.pyplot as plt
import numpy as np

# Replace these values with your actual scores
precision = [0.79, 0.73]
recall = [0.89, 0.56]
f1_score = [0.83, 0.63]

labels = ['Class 0', 'Class 1']

# Plotting the bar chart
width = 0.2
x = np.arange(len(labels))

fig, ax = plt.subplots()
rects1 = ax.bar(x - width, precision, width, label='Precision')
rects2 = ax.bar(x, recall, width, label='Recall')
rects3 = ax.bar(x + width, f1_score, width, label='F1-Score')

# Adding labels, title, and legend
ax.set_ylabel('Scores')
ax.set_title('Logistic Regression Model Evaluation Metrics')
ax.set_xticks(x)
ax.set_xticklabels(labels)
ax.legend()
```



6.2 Results:

```
The accuracy of the Logistic Regression model is: 0.7368421052631579
Classification Report:
              precision    recall  f1-score   support

           0       0.74      1.00      0.85        42
           1       0.00      0.00      0.00        15

    accuracy                           0.74        57
   macro avg       0.37      0.50      0.42        57
weighted avg       0.54      0.74      0.63        57
```

Comparision with different algorithms:

Logistic Regression is a supervised machine learning algorithm used for classification tasks, predicting the probability of an instance belonging to a given class. It's a statistical algorithm that analyses the relationship between a set of independent variables and the dependent binary variables. Other algorithms for classification tasks include Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbors, Support Vector Machines, and Neural Networks. Each has its own strengths and weaknesses, with some being easy to interpret (Decision Trees), reducing overfitting (Random Forest), useful for text classification (Naive Bayes), or powerful but computationally expensive (Neural Networks).

7.Conclusion:

This study attempts to analyse various supervised machine learning algorithms and select the most accurate model in detection of breast cancer. The work focused in advancement of predictive models with the help of python to achieve better accuracy in predicting correct outcomes. The analysis of result signifies that, integration of data, feature scaling along with different classification method and analysis provide markedly successful tool in prediction. It has also observed that the model misdiagnosed few patients with cancer when they were not having cancer and vice versa. Although, the model is accurate but when dealing with lives of people, further research in building the most accurate and precise model must be carried out for better performance of classification techniques and get the accuracy as close to 100% as possible. Thus, the tuning of each of the models is necessary with the building of more reliable model.

Achievements:

The project has created a tool using Python that can understand and make sense of complex Breast function data. This tool has been very useful in finding patterns and trends in the data that we didn't know about before. Additionally, through this project, we have gained valuable experience and knowledge in analyzing medical data using Python.

Future Work:

The next steps for this project include refining the analysis tool to incorporate more advanced machine learning algorithms for better predictive capabilities. Additionally, the team plans to extend the project to analyse other organ functions using similar methodologies. The ultimate goal is to create a comprehensive health analysis platform that can provide valuable insights into various aspects of human health.